Partial Translation of J. P. Application No. Hei 4 (1992)-32966 A

(11) J.P. Application No. Hei 4 (1992)-32966 A
(43) Publication Date: February 4, 1992
(54) Title of the Invention: DOCUMENT GENERATING DEVICE
(21) Application Number: Hei 2 (1990)-133590
(22) Application Date: May 23, 1990
(71) Applicant: RICOH CO LTD
(72) Inventor: ITO HIDEO

Translation of line 9 from the bottom of lower right column of page 456 (2) through line 14 of upper right column of page 457 (3)

The existent document is input to an input part 1 and the input document is divided into morphemes by a morpheme analytic part 2. Based on the results divided by the morpheme analytic part 2, the content of the existent dictionary 4 is adjusted in a dictionary adjustment part 3 so as to be adapted for the purpose of use.

The following is an explanation based on a specific example of the input sentence. For example, when 'A tooth aches me very much.' is input as the existent document to the input part 1, this input sentence is transmitted to the morpheme analytic part 2. The morpheme analytic part 2 divides the input sentence into 'A tooth / aches / me / very much.' Firstly, the morpheme list is retrieved as shown in Figure 3, thereby it is understood that the input sentence includes the following candidates of the combination of notation / part of speech (referred to as morpheme).

Table 1

と [to]	case particle
て [te]	connective particle
も [mo]	adverbial postpositional particle
とても [to-te-mo]	Adverb
むし歯 [mu-shi-ba]	Noun
が [ga]	case particle
痛む [i-ta-mu]	ma–row five–level conjugating type verb

At this time, only the word stem parts of the conjugating type words are registered in the morpheme list while ending parts are separately provided for each part of speech. When matching with the input, the concatenation of the word stem parts and the ending parts may be subjected to matching, thereby realizing the compression of the morpheme list and improving the retrieving efficiency.

Next, based on the table of the concatenation shown in Figure 4, the concatenation possibilities of all the morphemes are checked. In this case, with respect to only 'to-te-mo' (adverb), 'mu-shi-ba' (noun), 'ga' (case particle) and 'i-ta-mu' (ma–row five–level conjugating type verb), the concatenation is possible. If there are plural concatenation abilities, the most suitable concatenation covering the input is selected by utilizing information on easiness of concatenation between parts of speech, morpheme length, the kind of characters, the number of all clauses, the number of morphemes in the clauses, etc.

Thus, the results obtained by converting the input into rows of morphemes are transmitted to the dictionary adjustment part 3. The dictionary adjustment part 3 retrieves the corresponding entry in the existent dictionary 4 as shown in Figure 5 by utilizing the above-mentioned information for each morpheme. At the time of kana (Japanese syllabary)–kanji conversion, optional information used for selecting the entry (the entry is not necessarily limited to kanji and also includes hiragana such as particles, auxiliary verb, etc.) is adjusted so that the best conversion rate is obtained. For example, the frequency of use, the degree of concatenation, or the like, are increased by the unit amount.

Figure 4

	noun	Proper noun	Irregular noun	...
Noun	○	○	○	
Proper noun	○	○	×	
Personal noun	○	○	×	
Irregular conjugating noun of sa–row	○	○	○	
⋮				

In Figure 4, ○ means that the concatenation between the preceding part of speech and the following part of speech is possible; and × means that the concatenation between the preceding part of speech and the following part of speech is not possible.

Figure 5

Notation	Part of speech	Reading	Frequency of use
⋮	⋮	⋮	⋮
むし歯	Noun	mu-shi-ba	20
むすぶ	Ba–row five–level conjugating verb	mu-su-bu	10
⋮	⋮	⋮	⋮

(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **04032966 A**

(43) Date of publication of application: **04.02.92**

(51) Int. Cl **G06F 15/40**
G06F 15/20
G06F 15/38

(21) Application number: **02133590**

(22) Date of filing: **23.05.90**

(71) Applicant: **RICOH CO LTD**

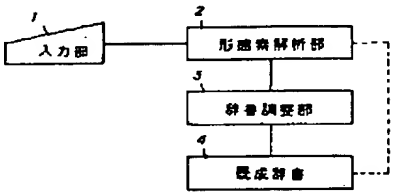
(72) Inventor: **ITO HIDEO**

(54) DICTIONARY GENERATING DEVICE

(57) Abstract:

PURPOSE: To perform a process in consideration of the unity of words by providing a dictionary adjustment part which adjusts the contents of a dictionary to the best contents for the purpose of its use according to the result of the division of an existent document which is inputted into morphemes.

CONSTITUTION: The existent document is inputted to an input part 1 and divided by a morpheme analytic part 2 into morphemes. For example, when 'A tooth aches me very much.' is inputted as the existent to the input part 1, the morpheme analytic part 2 divides the input sentence into 'A tooth/aches/me/very much.' Then the dictionary adjustment part 3 retrieves the corresponding entry in an existent dictionary 4 by utilizing information on easiness of concatenation between parts of speech, morpheme length, the kind of characters, the number of all clauses, the number of morphemes in the clauses, etc., by the morphemes and adjusts optional information which is used to select the entry at the time of KANA(Japanese syllabary)-KANJI(Chinese character) conversion so that the best conversion rate is obtained. Consequently, the utilization efficiency of the existent document is improved.



引用文献4

⑲日本国特許庁（JP）　⑪特許出願公開

⑫公開特許公報（A）　平4－32966

⑤Int. Cl.⁵		識別記号		庁内整理番号	㊸公開　平成4年(1992)2月4日
G 06 F	15/40	500	T	7056－5L	
	15/20	550	A	6914－5L	
	15/38		C	9194－5L	

審査請求　未請求　請求項の数 2　（全6頁）

㊹発明の名称　辞書作成装置

㉑特　願　平2－133590
㉒出　願　平2(1990)5月23日

㉒発明者　伊東　秀夫　東京都大田区中馬込1丁目3番6号　株式会社リコー内
㉑出願人　株式会社リコー　東京都大田区中馬込1丁目3番6号
㉔代理人　弁理士　高野　明近　外1名

明　細　書

1．発明の名称

辞書作成装置

2．特許請求の範囲

1．既存の文章を入力するための入力部と、文章を形態素に分割する形態素解析部と、既成の辞書と、前記入力部より入力された文章を前記形態素解析部によって形態素に分割した結果をもとに、前記辞書の内容をその使用目的に対して最適になるように調整する辞書調整部とを具備したことを特徴とする辞書作成装置。

2．既存の文章を入力するための入力部と、文章を形態素に分割する形態素解析部と、前記入力部より入力された文章を前記形態素解析部によって形態素に分割した結果をもとに辞書を自動作成する辞書作成部とを具備したことを特徴とする辞書作成装置。

3．発明の詳細な説明

技術分野

本発明は、辞書作成装置に関し、より詳細には、機械翻訳や仮名漢字変換に用いられる辞書作成装置に関する。

従来技術

一般に、自然言語処理においては、その対象言語の語彙情報の蓄積である辞書を必要とする。ところが以下のような問題がある。

まず、収集の問題としては、必要とされる語彙情報の範囲は明確でなく、十分とされる語彙情報の量は膨大であるため、語彙情報の収集は困難である。

次に、選択の問題としては、たとえ語彙情報が収集されたとしても、対象となる状況に対してどの語彙情報を使用すべきかを選択することは困難である。

さらに、管理の問題としては、たとえ語彙情報が収集されたとしても、その有効範囲や期間は不断に変化するため、語彙情報の有効性の維持管理は困難である。

このような問題に対処するために、十分な範囲を対象とせず、必要な範囲のみを対象として辞書

を作成又は調整することが重要となる。

従来においては、多大なコストをかけて上記語彙情報の収集の問題に対処し、ユーザの援助によって上記選択の問題に対処している（管理の問題はほとんど対処されていない）。しかし、この方法には多大なコストがかかることやユーザの負担が大きいことなどの問題がある。

また従来においては、例えば仮名漢字変換や機械翻訳における分野別辞書のように、対象とする領域を限定することで上記語彙情報の収集の問題及び選択の問題に対処している。しかし、この方法によって実用的な効果を得るためには、領域の限定範囲を極度に狭くすることが必要であり、したがって対象範囲がその限定範囲に含まれる保証がない場合、このような辞書を多数用意せねばならず、多大なコストがかかるという問題がある。また、どのように限定範囲を定めればよいかが明確でないという問題もある。また、そのような狭い限定範囲においては、上記管理の問題が発生しやすいという問題もある。

例えば、特開昭６０－１４７８６８号公報に記載されているものは、分野別又は個人用の既存の資料を利用して最適な漢字辞書を作成する辞書作成装置に関するもので、既存の文章に含まれる漢字部分の出現回数をカウントすることで、その文章又はそれに類する文章に対し、同音異義語の使用順位を調整した辞書を作成するものである。しかし、一般に自然言語は漢字のみで構成されるものではなく、漢字部分と漢字以外の部分を合わせた全体、もしくは、語としてのひとまとまりを考慮して処理を行わなくてはならない。

例えば、「むし歯」という単語が既存の文章中で使用されていた場合、この技術では漢字部分にのみ注目するため「むし歯」というひとまとまりでは処理されず、既成辞書中に「むし歯」が存在していても、読み「ム、シ、バ、ムシ、シバ、ムシバ」に対する同音異義語の使用順位を正しく調整することができない。また、漢字仮名変換辞書により漢字部分の「歯」の読みを得る場合でも、「むし歯」という１まとまりで処理されないため、「ハ」ではなく「バ」であるという結果を得るのは非常に困難である。

目　的

本発明は、上述のごとき実情に鑑みてなされたもので、辞書の調整や作成を自動的に行うことで収集の問題を解決し、対象領域に属する既存文章の分析結果を基に行うことで選択と管理の問題を解決し、既存文章の分析を漢字部分だけでなく全形態素に対して行うことで語としてのまとまりを考慮した処理を可能にするようにした辞書作成装置を提供することを目的としてなされたものである。

構　成

本発明は、上記目的を達成するために、（１）既存の文章を入力するための入力部と、文章を形態素に分割する形態素解析部と、既成の辞書と、前記入力部より入力された文章を前記形態素解析部によって形態素に分割した結果をもとに、前記辞書の内容をその使用目的に対して最適になるように調整する辞書調整部とを具備したこと、或いは、（２）既存の文章を入力するための入力部と、文章を形態素に分割する形態素解析部と、前記入力部より入力された文章を前記形態素解析部によって形態素に分割した結果をもとに辞書を自動作成する辞書作成部とを具備したことを特徴としたものである。以下、本発明の実施例に基づいて説明する。

第１図は、本発明による辞書作成装置の一実施例を説明するための構成図で、図中、１は入力部、２は形態素解析部、３は辞書調整部、４は既成辞書である。

既存の文章を入力部１に入力し、入力された文章を形態素解析部２により形態素に分割する。該形態素解析部２により形態素に分割された結果に基づいて既成辞書４の内容を使用目的に適合するように辞書調整部３で調整し、辞書を作成する。

以下に具体的な入力文に基づいて説明する。既成文章として「とてもむし歯が痛む」が入力部１に入力された場合、この入力文は形態素解析部２に送られる。形態素解析部２は、次のようにして

前記入力文を「とても／むし歯／が／痛む」という形態素に分割する。まず、第３図に示すような形態素リストを検索し、入力中に下記の第１表のような表記／品詞の組（形態素と呼ぶ）の候補が含まれることがわかる。

第１表

と	格助詞
て	接続助詞
も	副助詞
とても	副詞
むし歯	名詞
が	格助詞
痛む	ま行五段動詞

この際、活用語は語幹部分のみ形態素リストに登録し、一方活用語尾を品詞毎に別に持ち、入力とのマッチングの際には、これらを連接したものをマッチング対象とすることで、形態素リストの圧縮と検索効率向上を図ってもよい。

次に、第４図に示すような品詞間の連接表をもとに、上記形態素の連接可能性を全て調べる。この場合は

とても（副詞）　むし歯（名詞）

が（格助詞）　痛む（ま行五段動詞）

という連接のみが可能であるが、複数の連接可能性がある場合は、品詞間の連接しやすさ、形態素長、字種、全体の文節数、文節に含まれる形態素数などの情報を用いて入力を覆う最も適当な連接の組を選択する。

このようにして、入力を形態素の列に変換した結果は辞書調整部３に送られる。辞書調整部３は、上記形態素毎に表記の情報を利用し、第５図に示すような既成辞書４中の該当エントリを検索する。そして、仮名漢字変換においてエントリ（漢字とは限らない。助詞、助動詞など平仮名のエントリも含む）の選択に使用される任意の情報を、変換率が最適になるように調節する。例えば、使用度や連接度などを単位量だけ増加させる。

ここで使用度とは例えば使用頻度などそのエントリ単独に関する使用されやすさを表す測度であり、連接度とはそのエントリが前後のどのような種類のエントリと連接しやすいかを表す測度である。

以上のようにして、既存の文章の形態素解析の結果をもとに、既存辞書を仮名漢字変換率を最適にするように調節することができる。

なお、第１図に点数で示したように、形態素解析部が使用する形態素リストは既成の辞書で兼用してもよい。すなわち、仮名漢字変換のための品詞分類と形態素解析のための品詞分類を同一にし、第４図に示す連接表は、その品詞分類に対して作成されていればよい。

第６図は、第１図に基づく分類による辞書作成装置の動作を説明するためのフローチャートである。以下、各ステップに従って順次説明する。

step1：既存の文章を入力する。

step2：入力された文章を形態素リストを検索し、表記を品詞の組の候補を生成する。

step3：品詞間の連接表をもとに連接可能性を調べる。

step4：最も適当な連接の組を決定する。

step5：連接の組を構成する形態素を１つ抽出する。

step6：既成辞書の該当エントリを検索する。

step7：エントリに関する情報を調整する。

step8：最後の形態素かどうか判断する。最後の形態素であれば終了し、そうでなければstep5に戻る。

第２図は、辞書調整部を有しない辞書作成装置の構成図で、図中、５は入力部、６は形態素解析部、７は辞書作成部である。すなわち、既成の辞書を調節するのではなく、新たに仮名漢字変換用辞書を作成するためには、次のようにすればよい。既存の文章を入力部５に入力し、入力された文章を形態素解析部６により形態素の列に分割する。ただし、形態素解析部６で使用する品詞分類は、目的とする仮名漢字変換で使用するものと同一もしくは対応づけが可能なものとする。

次に、辞書作成部７は形態素の読み（平仮名）を得るために、非平仮名－平仮名変換を行う。非平仮名－平仮名変換（今後単に仮名変換と呼ぶ）とは、漢字・カタカナ・数字・記号などが入力中にあると、それに対応する読みを平仮名列として出力するものである。

例えば、「平成４年のバルセロナ」という入力に対して形態素解析部をした結果、以下の第２表に示すような形態素列が得られた場合、仮名変換の結果は第３表のように各形態素に対する読みが平仮名で与えられたものが出力される。

第２表

平成	元号
4	数字
年	助数詞
の	連体助詞
バルセロナ	固有名詞

第３表

平成	元号	へいせい
4	数	よ
年	助数詞	ねん
の	連体助詞	の
バルセロナ	固有名詞	ばるせろな

前記仮名変換は、各形態素の表記またはその一部をキーとして、第５図のような既存の仮名漢字変換辞書、または第７図のようなカタカナ・記号・数字の読み対応表、または第８図のような単漢字音訓表、または第９図のような漢字仮名変換表を検索して対応する読みを得ることで行う。

次に、辞書作成部７は、各形態素が入力の既存文章中に何回現れたかをその使用度としてカウントする。

最後に、辞書作成部７は、以上までで得られた形態素毎の表記、品詞、読み、使用度を、第５図のような仮名漢字変換辞書の形式にして出力する。

また、辞書を作成する場合、まったく新規に作成するのではなく、一般に使用頻度が高い日常語や基本語の辞書を共通辞書として用意し、入力文章に対し辞書を作成する場合は、その共通の辞書にマージするようにしてもよい。

あるいは、同種類の入力文章を数回に分けて入力する場合、旧辞書に含まれない形態素に対しては単にエントリとして加え、旧辞書にすでに含まれている形態素に対しては旧辞書中の使用度と今回カウントされた使用度を加算するようにして、旧辞書を成長させるようにしてもよい。

第１０図は、第２図に基づく本発明による辞書作成装置の動作を説明するためのフローチャートである。以下、各ステップに従って順次説明する

step1：既存の文章を入力する。

step2：入力された文章を形態素リストを検索し、表記と品詞の組の候補を生成する。

step3：品詞間の連接表をもとに連接可能性を調べる。

step4：最も適当な連接の組を決定する。

step5：連接の組を構成する形態素を１つ抽出する。

step6：前記形態素に関し、非平仮名・平仮名変換を行い「読み」の情報を得る。

step7：形態素の出現回数を基に使用度を求める。

step8：前記step7までで得られた情報を形態素毎に記憶装置に保存する。

step9：最終の形態素かどうか判断する。最後の形態素でなければstep5に戻る。

step10：step9において最後の形態素であれば前記step8までで得られた情報を辞書の形式にして出力する。

本発明の実施例では仮名漢字変換辞書の調整や作成について説明したが、機械翻訳用辞書、漢字仮名変換辞書など、辞書を具備した自然言語処理装置ならば、適宜変形して実施することができる。また、既存文章の入力方式、形態素解析の方式、辞書内容と検索方式に関して限定するものではない。要するに、既存文章を利用し、それを形態素に分割し、その情報を利用して辞書を調整、作成する場合に適用できる。

効　果

以上の説明から明らかなように、本発明によると、辞書の調整や作成を自動的に行うことで、対象言語の語彙情報の蓄積である辞書に関する収集の問題、選択の問題、管理の問題を解決することができるとともに、既存文章の利用効率を漢字部分のみ利用する場合に比較して大幅に向上することができる。

４．図面の簡単な説明

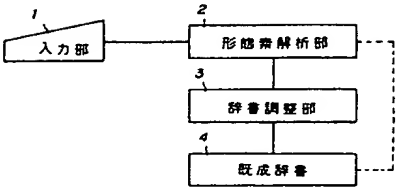
第１図は、本発明による辞書作成装置の一実施例を説明するための構成図、第２図は、辞書調整部を有しない辞書作成部の例を示す構成図、第３図は、表記と品詞と示す図、第４図は、品詞間の連接を示す図、第５図は、既成辞書の内容を示す

図、第６図は、第１図に基づく辞書作成装置の動作を説明するためのフローチャート、第７図は、カタカナ・記号・数字の読みの対応を示す図、第８図は、単漢字の音訓を示す図、第９図は、漢字仮名変換を示す図、第１０図は、第２図に基づく辞書作成装置の動作を説明するためのフローチャートである。

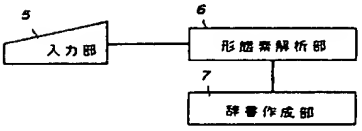
１…入力部、２…形態素解析部、３…辞書調整部、４…既成辞書。

特許出願人　株式会社　リコー
代理人　高野　明近
（ほか１名）

第１図



第２図



第３図

表記	品詞
： むし曲 むすぶ ：	： 名詞 ぱ行五段動詞 ：

第４図

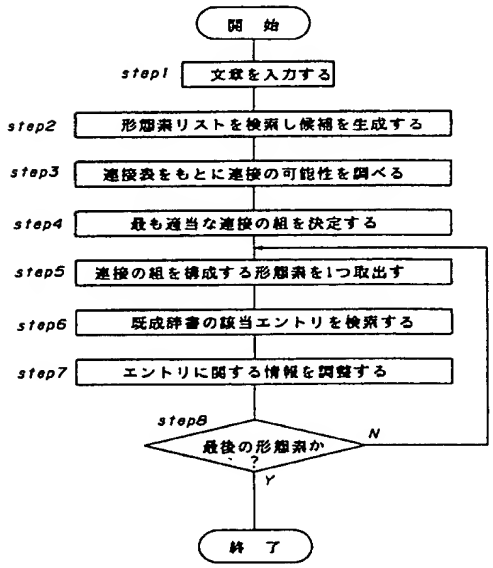
前＼後	名詞	固有名詞	サ変名詞	・・・
名詞	○	○	○	
固有名詞	○	○	×	
人称名詞	○	○	×	
サ変名詞	○	○	○	
：				

上記において、○は前の品詞と後の品詞が連接可能なことを示し×は連接できないことを示す。

第５図

表記	品詞	読み	使用度
： むし曲 むすぶ ：	： 名詞 ぱ行五段動詞 ：	： むしば むすぶ ：	： 20 10 ：

第６図



第 7 図

非平仮名	読み
ア	あ
イ	い
ウ	う
・	・
・	・
#	しゃーぷ
・	・

第 8 図

漢字	読み
・	・
・	・
歯	音)は 音)ば 訓)なし
・	・
・	・

第 9 図

単語	読み
・	・
・	・
むし歯	むしば
・	・
・	・

第 10 図